O'Brien et al. (2014) accepted for publication in Comp. Stat. and Data. Anal. on 2 June 2014.

Reducing the computational cost of the ECF using a nuFFT: A fast and objective probability density estimation method

TRAVIS A. O'BRIEN*AND WILLIAM D. COLLINS†

Lawrence Berkeley National Lab, Berkeley, CA, USA

†*and University of California, Berkeley, CA, USA*

SARA A. RAUSCHER AND TODD D. RINGLER

Los Alamos National Lab, Los Alamos, NM, USA

Abstract

A nonuniform, fast Fourier transform can be used to reduce the computational cost of the empirical characteristic function (ECF) by a factor of 100. This fast ECF calculation method is applied to a new, objective, and robust method for estimating the probability distribution of univariate data, which effectively modulates and filters the ECF of a dataset in a way that yields an optimal estimate of the (Fourier transformed) underlying distribution. This improvement in computational efficiency is leveraged to estimate probability densities from a large ensemble of atmospheric velocity increments (gradients), with the purpose of characterizing the statistical and fractal properties of the velocity field. It is shown that the distribution of velocity increments depends on location in an atmospheric model and that the increments are clearly not normally distributed. The estimated increment distributions exhibit self-similar and distinctly multifractal behavior, as shown by structure functions that exhibit power-law scaling with a non-linear dependence of the power-law exponent on the structure function order.

1 Introduction

Research often calls for the estimation of probability distribution functions (PDFs) derived from empirical data. For instance, information about a distribution may be necessary to assess whether differences between two sets of data are statistically significant, or it may be required to estimate the probabilities that outliers come from the distribution of a given dataset. A variety of PDF approximations (e.g., histograms) are frequently used to represent the relative occurrence of data values.

This paper describes a computationally efficient method to estimate probability distributions based on recent work by Bernacchia and Pigolotti [1]. We have developed this technique to support research on scaling in the Earth's atmosphere, but the method should be generally applicable across the physical and engineering disciplines. We have initially applied this method to aid the development of a theory about resolution dependence in atmospheric models. The following discussion necessarily makes heavy use of some terms that are commonly used in the atmospheric sciences but that may be unfamiliar to researchers from other fields. Appendix A provides definitions for some of these terms.

Many studies also require strictly non-parametric estimation procedures so that the resulting PDFs are free of *a priori* assumptions regarding their underlying functional forms. In our particular application, the normality of velocity gradients is a key hypothesis that should be proven or disproven from the emergent properties of the data itself without recourse to Gaussian fitting. The traditional methods for estimating PDFs, e.g. binning methods and kernel density techniques, require specification of a bandwidth parameter that heavily influences the shape of the resulting PDF [2, 3, 4]. While methods exist for estimating an optimal bandwidth, these methods usually require some assumption about the shape of the underlying PDF [5, 4, 1]. Given that our application requires an unbiased determination of the normality of the velocity increments, estimation methods utilizing such assumptions would not be suitable for our analysis. While methods do exist for testing normality (e.g., [6]), our analysis additionally requires that we estimate various moments of the distribution. One can readily and efficiently perform tests for normality and estimate moments if an estimate of the underlying distribution is available.

For these reasons, the method of Bernacchia and Pigolotti [1] for estimating PDF distributions should in principle be well suited for such an application because it provides an objective PDF estimate that requires no prior assumptions regarding the underlying distribution. Bernacchia and Pigolotti [1] derive an expression for a data-derived, optimal kernel [7] and the resulting and self-consistent kernel density estimate; their kernel derivation method is even optimal for multi-modal data. This 'self-consistent' estimate converges on the true distribution at a faster rate than traditional binning or kernel density estimation methods

[1].

However, during our initial attempts to apply the PDF estimation method of Bernacchia and Pigolotti [1] (hereafter BP11), we discovered that its computational performance is not practicable. Most of the BP11 algorithm is implemented in inverse Fourier space and is based on transforming the data under analysis to inverse Fourier space by computing its empirical characteristic function, $\mathscr{C}_n$, given by:

$$\mathscr{C}_n = \sum_{j=1}^{N} e^{i\cdot\chi_j\cdot\tau_n}, \tag{1}$$

where χ_j are a collection of data points that are presumed to come from a random distribution, N is the number of data points, and τ_n are the frequencies at which the empirical characteristic function is calculated. Calculation of $\mathscr{C}_n$ is equivalent to an inverse discrete Fourier transform in which the Fourier coefficients are $a_j = 1$ for each of the χ_j data points. Since the direct calculation of the discrete Fourier transform is notoriously slow, it would be preferable to evaluate this discrete Fourier transform using the fast Fourier transform (FFT) method of Cooley and Tukey [8]. However, the FFT is not directly applicable since it requires that the Fourier coefficients are specified on an evenly-spaced grid. This requirement is violated since the χ_j data points are presumed to be randomly distributed, and so their spacings are also random.

In this paper, we show how to accelerate the computational performance of the BP11 density estimation method using the nonuniform FFT (nuFFT) method of Greengard and Lee [9] to approximate the empirical characteristic function (Section 2). We demonstrate that this method substantially improves the speed of the BP11 density estimation method without compromising its accuracy or convergence properties (Section 3). We apply this method to estimate the PDF of velocity increments from atmospheric model output in support of a hypothesis relating velocity increments to model resolution dependence (Section 4). We show that the increments from a specific atmospheric model are generally bell-shaped but demonstrably non-Gaussian. Further, we use the estimated distributions to show that the velocity field is self-similar and multifractal. This ability to rapidly characterize increment distributions has thus proved invaluable in our development of a robust theory on resolution dependence in atmospheric models.

2 Estimating the self-consistent density via FFT

2.1 Summary of the BP11 self-consistent density estimation method

Kernel density estimation is a widely used method for estimating the probability distribution function (PDF) of a given dataset (e.g., [2, 3]), in which the PDF is approximated as a normalized sum of *kernel* functions $K(\chi)$ centered on each data point χ_j:

$$f^{KDE}(\chi) = \frac{1}{N}\sum_{j=1}^{N} K(\chi-\chi_j).$$

The choice of $K(\chi)$–particularly the width of K–can heavily influence f^{KDE}, and there is a host of literature devoted to choosing the kernel width. Except in some specific circumstances (e.g., the data are known to be normally distributed [2]), the choice of the kernel and the kernel width are subjective [10, 1, 4]. Bernacchia and Pigolotti [1] recently derived a method for objectively estimating the probability distribution function (PDF) of a univariate dataset. They show that the dataset itself can be used to derive a kernel (both its shape and width) in an objective, data-driven way. We summarize the essential details of the derivation and the method here.

The inverse Fourier transform of the KDE estimate is simply the product of the transform kernel and the ECF of the data; we derive this relationship here, since it is relevant for understanding the role of the nuFFT in the BP11 method. Recognizing that a kernel density estimate is equivalent to a sum of convolutions between a kernel function and delta functions centered on the data:

$$\begin{aligned} f^{KDE}(\chi) &= \frac{1}{N}\sum_{j=1}^{N} K(\chi-\chi_j) \\ &= \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty} K(s)\cdot\delta(\chi-\chi_j-s)\,ds \\ &= \frac{1}{N}\sum_{j=1}^{N} K(\chi)*\delta(\chi-\chi_j), \end{aligned}$$

the kernel density estimate can readily be transformed to its inverse Fourier-space representation, ϕ^{KDE} using the convolution theorem:

$$\begin{aligned} \phi^{KDE}(\tau) &= \mathscr{F}_\tau{}^{-1}\left[f^{KDE}\right] \\ &= \mathscr{F}_\tau{}^{-1}\left[\frac{1}{N}\sum_{j=1}^{N} K(\chi)*\delta(\chi-\chi_j)\right] \\ &= \kappa(\tau)\cdot\frac{1}{N}\sum_{j=1}^{N} e^{i\chi_j\tau} \\ &= \kappa(\tau)\cdot\mathscr{C}(\tau), \end{aligned}$$

where $\mathscr{F}_\tau{}^{-1}$ represents the inverse Fourier transform from data space, χ, to inverse Fourier space, τ; κ represents the inverse Fourier transform of K; and $\mathscr{C}$ represents the empirical characteristic function of the data.

Bernacchia and Pigolotti [1] use this relationship and the result of Watson and Leadbetter [7], which states that the mean squared error of a kernel density estimate is minimized if the kernel satisfies the equation: $\hat{\kappa} = N\cdot(N-1+|\phi|^{-2})^{-1}$. They use this optimal kernel to provide an equation for the optimal PDF estimate (in inverse Fourier space):

$$\hat{\phi}(\tau) = \hat{\kappa}(\tau)\cdot\mathscr{C}(\tau) = \mathscr{C}(\tau)\cdot\frac{N}{N-1+|\phi|^{-2}}. \tag{2}$$

Since the underlying distribution (and its transform, ϕ) is assumed to be unknown, they derive a solution to Equation 2 using an iterative procedure in which an initial guess at ϕ, ϕ_0, is used to estimate $\hat{\phi}_1$ which is then used as the next guess at ϕ to estimate $\hat{\phi}_2$, and so on. They show that if this iterative procedure converges, that it will converge to a solution $\phi^{sc} = \kappa^{sc} \cdot \mathscr{C}$ (which provides a self-consistent solution to Eqn 2: $\phi^{sc}(\tau) = \mathscr{C}(\tau) \cdot N \cdot [N-1+|\phi^{sc}(\tau)|^{-2}]^{-1}$), provided κ^{sc} satisfies the following equation, which is a function of the ECF amplitude:

$$\kappa^{sc}(\tau) = \frac{N}{2(N-1)} \left[1 + \sqrt{1 - \frac{4(N-1)}{N^2 |\mathscr{C}(\tau)|^2}} \right] I_A(\tau), \qquad (3)$$

where $I_A(\tau)$ represents a frequency filter that is 1 for the set of accepted frequencies A (defined below), and 0 otherwise.

In order for Equation 3 to provide a stable solution to Equation 2, the set of accepted frequencies must be specified such that $|\mathscr{C}(\tau)|^2 \geq 4(N-1)N^{-2}$ for $\tau \in A$. Further, the frequency set A may exclude an arbitrary additional subset of otherwise acceptable frequencies, which reflects the arbitrariness of the initial guess ϕ_0 of the iterative solution. Bernacchia and Pigolotti [1] show that ϕ^{sc} converges to the true underlying distribution as N increases, provided that a number of conditions are met, including integrability of the characteristic function and boundedness of A. The stability condition on A forces $\kappa^{sc}(\tau)$ to be real-valued, implying that its data space representation $K^{sc}(\chi)$ is symmetric.

Finally, this self-consistent estimate can be Fourier transformed to obtain the data-space estimate of the PDF: $f^{SC}(\chi) = \mathscr{F}_\chi [\kappa^{sc}(\tau) \cdot \mathscr{C}(\tau)]$. Provided that the ECF has been calculated, calculation of $\kappa^{sc}(\tau)$ is trivial, so the bulk of the cost of computing $f^{SC}(\tau)$ comes from the computation of the ECF.

2.2 *Reducing the computational cost of the ECF using a nuFFT*

While exploring this BP11 density estimation method, it became clear that the ECF itself is a type of direct Fourier transform (DFT):

$$\mathscr{C}(\tau) \propto \sum_{j=1}^{N} a_j \cdot e^{i\chi_j \tau},$$

where χ_j represents abscissa values in data space, τ represents abscissa values in inverse Fourier space, and the a_j Fourier coefficients are all 1. Since the χ_j values are assumed to be randomly distributed, they presumably are not regularly spaced, which excludes the possibility of using a standard FFT method to evaluate the DFT. However, the nonuniform FFT (nuFFT) method described by Greengard and Lee [9] is specifically designed to reduce the computational cost of DFTs on irregularly-spaced data. The nuFFT method can be summarized as follows.

An arbitrary dataset of abscissa and ordinate pairs, χ_j and a_j, can be viewed as a continuous function that is a sum of weighted delta functions:

$$a(\chi) = \sum_{j=1}^{N} a_j \cdot \delta(\chi - \chi_j).$$

Convolution of $a(\chi)$ with a Gaussian g_h spreads the delta functions across the abscissa, which results in a smooth curve: $a'(\chi) = a(\chi) * g_h(\chi)$. By the convolution theorem, the Fourier transform of $a'(\chi)$, $c'(\tau)$, is proportional to the Fourier transform of $a(\chi)$, $c(\tau)$:

$$\begin{aligned} c'(\tau) &= \mathscr{F}_\tau\left(a'(\chi)\right) \\ &= \mathscr{F}_\tau\left(a(\chi) * g_h(\chi)\right) \\ &= c(\tau) \cdot \tilde{g}_h(\tau), \end{aligned}$$

where $\tilde{g}_h(\tau)$ is the Fourier transform of g_h.

If the abscissa is sampled at regular intervals, χ_k, then a FFT technique can readily be used to approximate the Fourier transform of $a'(\chi_k)$. Finally, the convolution theorem is used to deconvolve $c'(\tau_n)$ (divide c' by $\tilde{g}_h$), which results in an approximation of the discrete Fourier transform of the irregularly-spaced (χ_j, a_j) data. Greengard and Lee [9] show that the nuFFT can approximate the DFT with arbitrary accuracy, which is controlled by the interaction of three main factors: the width h of the convolving Gaussian; the number surrounding χ_k values at which the convolution is calculated for each (χ_j, a_j) point; and the spacing of the χ_k grid. The speed of the nuFFT method, which is a trade-off for accuracy, is also controlled by these three factors.

With respect to using the nuFFT to calculate the ECF, the (χ_j, a_j) abscissa/ordinate pairs are identically $(\chi_j, 1)$, where χ_j represent the random (irregularly spaced) data. With all the a_j values set to 1, the convolution step effectively reduces to a (unnormalized) kernel density estimate of the data:

$$a'(\chi) = \sum_{j=1}^{N} g_h(\chi - \chi_j).$$

So in statistical terms, the essential steps of the nuFFT approximation of the ECF can be summarized as: (1) perform a kernel density estimate (on a regular grid), (2) use an inverse FFT to transform the kernel density estimate to inverse Fourier space, and (3) divide the transformed density by the inverse Fourier transform of the kernel function.

2.3 *A Fast BP11 algorithm*

The following steps summarize the algorithm that we use to perform a fast and efficient calculation of the BP11 density estimate (for conciseness, we hereon express functions at a given grid point using the function symbol and the corresponding grid subscript: e.g., $\mathscr{C}_n \equiv \mathscr{C}(\tau_n)$):

1. Configure a regular grid and its transform grid: χ_k and τ_n.
2. Specify a Gaussian kernel $g_h(x) = exp\left[-(x/h)^2\right]$.
3. Convolve the data with the Gaussian to obtain a (unnormalized) kernel density estimate:

$$f'_k = \frac{1}{N} \sum_{j=1}^{N} \chi_j \cdot g_h\left(\frac{\chi_k - \chi_j}{h}\right).$$

4. Perform an inverse FFT of the kernel density estimate to obtain its transform: $\phi'_n = \mathscr{F}^{-1}_{\tau_n}(f'_k)$.

5. Divide ϕ'_n by the transform of the Gaussian kernel to deconvolve the FFT and obtain an estimate of the empirical characteristic function: $\mathscr{C}_n \approx \phi'_n \cdot [\tilde{g}(\tau_n)]^{-1}$.

6. Calculate the self-consistent kernel transform κ^{sc}_n (Eqn 3).

7. Calculate the self-consistent PDF transform: $\phi^{SC}_n = \kappa^{sc}_n \cdot \mathscr{C}_n$.

8. Perform an FFT to obtain the self-consistent PDF estimate: $f^{sc}_k = \mathscr{F}(\phi^{sc}_n)$.

If applied naïvely, the convolution in step (3) can be as expensive as the direct DFT calculation (or more so). For N data points, a full calculation of the convolution requires $\mathscr{O}(N^2)$ calculations, whereas the direct DFT calculation requires $\mathscr{O}(N \cdot M)$ calculations for M frequency points and hence would be faster if $M < N$. The speed of the convolution can be dramatically improved if the Gaussian contribution from each of the χ_j data points is only applied to a limited set of q surrounding points. To this end, Dutt and Rokhlin [11] provide an expression for specifying the width of the Gaussian h and the point-width q of the convolution such that the resulting FFT is the same as the direct DFT within a specifiable accuracy. The convolution part of this algorithm requires $\mathscr{O}(N \cdot q)$ calculations and the FFT portion requires $\mathscr{O}(M \cdot \log M)$ [8]. Simple algebraic manipulation can show that if $q < M$ and $\log M \ll N$, then $N \cdot q + M \cdot \log M < N \cdot M$, and so the nuFFT is theoretically faster than the direct DFT calculation. These conditions also imply that the nuFFT-based calculation is theoretically $\mathscr{O}(M/q)$ times faster than the direct calculation.

To simplify the analysis of velocity increment PDFs and to provide a static grid on which all of the estimated PDFs can be stored, we standardize the data (i.e., $\chi_j = (\chi'_j - \bar{\chi}) \cdot \sigma^{-1}_\chi$) prior to applying the density estimation algorithm ($\bar{\chi}$ and σ_χ are the mean and standard deviation of the original χ'_j data respectively). We specify χ_k as 4,097 evenly-spaced points from -20 to 20 unit standard deviations. Since in our analysis the χ_j data points are all real, the Fourier transform of these points has Hermitian symmetry and hence the redundant negative frequency components of the transform may be ignored. Therefore the χ_k grid yields a transform grid τ_n with 2,049 evenly-spaced frequency points. We only consider the lowest half of the frequency points (i.e., we set $\mathscr{C}_n = 0$ for $n >$ 1,025) since the nonuniform FFT method is only guaranteed to provide a good approximation over this range [11]. Following Dutt and Rokhlin [11], we specify the width of the convolution kernel as h=1.5629, and we apply the convolution to the q=28 χ_k nearest points surrounding each χ_j data value. We find that this configuration produces an approximation of the ECF that differs from the exact DFT calculation by less than 10^{-7} over all considered frequencies (see Section 3 and Figure 1).

We also note that we implemented the selective frequency filter, I_n, in a slightly different manner than Bernacchia and Pigolotti [1]. They show that the self-consistent density estimate converges on the true density provided the filter I_n is set to 1 for some subset of the frequencies for which $\mathscr{C}$ is above the estimate stability threshold given by $|\mathscr{C}_n|^2 \geq 4 \cdot (N-1) \cdot N^{-2}$ and set to 0 for all other frequencies. Whereas they choose the subset based on a frequency cut-off t^* such that $\mathscr{C}$ is above the stability threshold for half of the frequencies within $[-t^*, t^*]$, we choose a cut-off frequency based on the occurrence of three consecutive $\mathscr{C}$ values below the stability threshold. In our implementation $I_n = 0$ for all $n > n^*$, where n^* is the index of the lowest frequency for which $\mathscr{C}_{n^*+1}$, $\mathscr{C}_{n^*+2}$, and $\mathscr{C}_{n^*+3}$ are below the stability threshold. We choose this criterion because it is fast to implement and we find that it avoids an occasional, spurious leakage of high-frequency components that manifests as high-frequency waves superimposed on the density estimate. Bernacchia and Pigolotti [1] note that the selection of the subset of frequencies is arbitrary and corresponds to the arbitrary choice of initial density estimate in the iterative procedure that they use to derive the expression for $\hat{\phi}$. As long as the subset is bounded and the bound grows with N, a self-consistent estimate will converge. Our filter choice satisfies these criteria for integrable characteristic functions, since the stability threshold decreases with increasing N and therefore higher frequencies are permitted as N increases. Therefore our implementation of the I_n filter maintains the convergence properties of the BP11 density estimate (see Section 3 for verification of this).

3 Evaluating against artificial data

3.1 N^{-1} convergence for the nuFFT-based method

To show that the FFT-based approximation of the empirical characteristic function $\mathscr{C}^{(FFT)}$ reproduces the exact and direct calculation $\mathscr{C}^{(DFT)}$ at high precision, we compare the two quantities calculated from the samples drawn from a normal distribution with sample sizes ranging from 64 to 4,096. Figure 1a shows the absolute difference between the two quantities, $|\mathscr{C}^{(FFT)} - \mathscr{C}^{(DFT)}|$, as a function of frequency for several different sample sizes. The FFT-based estimate differs from the true estimate by less than 10^{-7} or less over the entire frequency range. For reference, the inset of Figure 1 shows $\mathscr{C}^{(FFT)}$.

Because the FFT-based approximation of the ECF differs from the true calculation of the ECF by such a small amount, the convergence properties of the BP11 density estimate are unaffected. Figure 1b shows the mean squared error, $E_2(N) = \sum^N_j |\mathscr{N}(\chi_i) - \hat{f}(\chi_i)| \cdot \Delta\chi$, where $\mathscr{N}(\chi)$ is the normal distribution and $\Delta\chi$ is the grid spacing. $E_2(N)$ declines following N^{-1} for sample sizes ranging from 2^1 to 2^{19} in agreement with the convergence rate presented by BP11. While the convergence-rate of the FFT-based method is in accord with the convergence rate from BP11 over the range of sample sizes shown, the FFT-based method should have a lower-bound on E_2 that is controlled by the approximation error, $\varepsilon = |\mathscr{C}^{(FFT)} - \mathscr{C}^{(DFT)}| \sim 10^{-7}$. If the approximation-error of the density estimate, $|\hat{f}^{(FFT)} - \hat{f}^{(DFT)}|$, is larger than the nuFFT

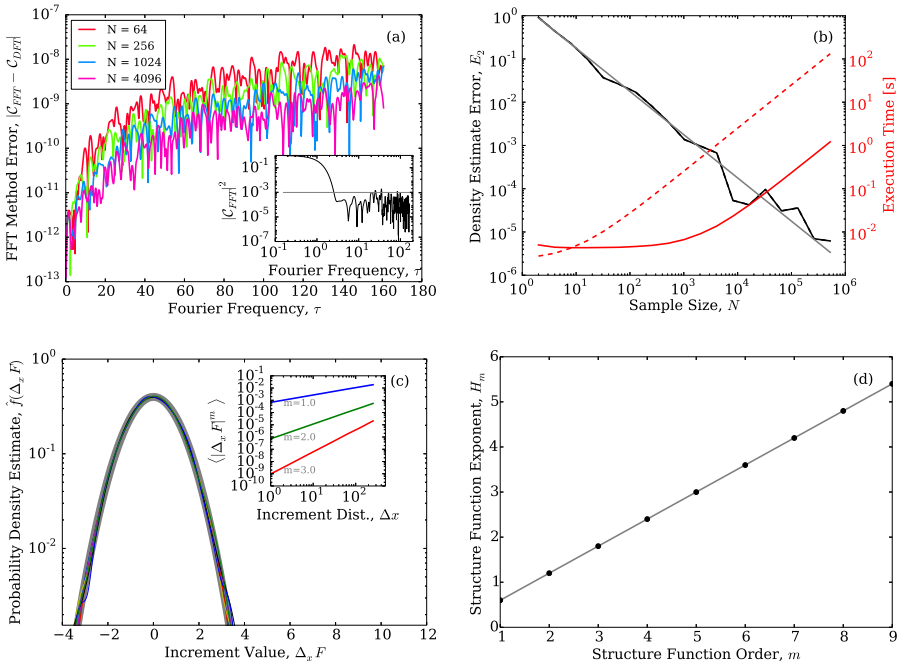

Figure 1: (a) The absolute difference between the FFT approximation of $\mathscr{C}_n$ and its exact value calculated via DFT. The inset shows the squared magnitude of the FFT-based estimate of $\mathscr{C}_n$. (b) The convergence of the average absolute error squared, and the calculation time as a function of sample size. The black line shows the calculated error from samples drawn from a normal distribution, the grey line shows N^{-1} convergence, the solid red line shows the calculation time from the FFT-based estimate, and the dashed red line shows the calculation time from the direct calculation of $\mathscr{C}_n$. (c) The (normalized) probability distributions of increments, at various distances, from an fBm field with H=0.6 on a semilog plot. The inset shows the first–third order structure functions from these distributions, which should vary as a power law for a self-similar field. Please refer to A for definitions of terms. (d) The power-law exponents from the mth order structure functions. The grey line shows $H_m = H \cdot m$, which is expected for an fBm field.

approximation error in the ECF, then E_2 will be dominated by ε, so E_2 will have a lower bound of $E_2 \sim \mathscr{O}(\varepsilon^2) \sim \mathscr{O}(10^{-14})$. In this case, E_2 for the FFT-based method should flatten out for sample sizes larger than $N \sim \mathscr{O}(10^{14})$, since $E_2 \propto N^{-1}$. This non-convergence for extremely large sample sizes could be mitigated by increasing the width of the Gaussian kernel (both h and q) to achieve a more precise estimate of the DFT. For the analysis in this manuscript, however, the sample sizes will not be so large that E_2 approaches its limit.

Figure 1b also shows the time required to perform the density estimates from both the FFT-based method and the DFT method. As described in Section 2, the FFT-based method scales as $\mathscr{O}(N \cdot q + M \cdot \log M)$ whereas the direct method scales as $\mathscr{O}(N \cdot M)$. Because we use $M = 2{,}049$ for both methods and for all sample sizes, both methods scale proportionally to N^1 for large sample sizes, as evinced by the parallel lines in Figure 1b. For sample sizes larger than $\mathscr{O}(10^3)$, the FFT-based method is approximately 100 times faster than the DFT method. For an $\mathscr{O}(10^6)$ sample size, the FFT method takes $\mathscr{O}(1)$ second versus the $\mathscr{O}(10^2)$ seconds for the DFT calculation.

3.2 *Increment PDF estimates from an fBm field*

In anticipation of the analysis presented in Section 4, we show a sample version of the same analysis applied to a dataset with well-known properties that mimics the data to which this method is applied in Section 4. The analysis in Section 4 has two goals: (1) to determine whether velocity increments are distributed normally, and (2) to show that the width and moments of the increment PDFs scale as a power-law of increment distance (as expected for a field with self-similar, fractal behavior).

Because atmospheric velocities are known to exhibit statistical self-similarity in reality and in models [12, 13, 14], we apply the analysis to a realization of a fractional Brownian motion, which is a type of self-similar field [15]. Fractional Brownian motion (fBm) can be categorized as a type of 'red-noise' field where the power spectrum of the field decays following a power-law: i.e., $P(f) \sim f^{-\beta}$, where P is the spectral power of the fBm field, f is the Fourier frequency, and β is the scaling exponent. fBm fields are characterized by their Hurst parameter H [15], which is directly related to β for fBm fields by the relationship

$H = (\beta - 1)/2$ [16].

Davis et al. [16] define the mth-order structure function of a field F as $S_m^F(\Delta x) \equiv \langle |F(x) - F(x+\Delta x)|^m \rangle$, which is the mth (absolute) moment of the PDF of increments calculated at distance Δx (we define $\langle \ldots \rangle$ as the average). For fBm fields, the structure functions scale as a power-law of increment distance: $S_m^F(\Delta x) \sim \Delta x^{H_m}$. The exponent for the mth-order structure function H_m is simply related to the Hurst exponent of the fBm field by $H_m = H \cdot m$ [17]. If we define the increment as $\Delta_x F \equiv F(x) - F(x+\Delta x)$, then the mth-order structure function can be calculated from the increment PDF by

$$S_m^F(\Delta x) = \int_{-\infty}^{\infty} |\Delta_x F|^m \cdot P(\Delta_x F)\, d\Delta_x F \sim \Delta x^{H \cdot m}. \quad (4)$$

Since the fBm field is generated based on samples drawn from a normal distribution, it can be shown that the distribution of increments are also normally distributed [17]. Therefore we expect $P(\Delta_x F)$ to be a normal distribution with variance $S_2^F(\Delta x) \sim \Delta x^{2H}$, implying that

$$P(\Delta_x F) = \frac{1}{\sigma_o \Delta x^H \sqrt{2\pi}} \cdot e^{-(\Delta_x F)^2 / 2(\sigma_o \Delta x^H)^2}, \quad (5)$$

where σ_o is a constant of proportionality related to the total variance of the field F. This is the form of the PDF for any self-similar field with increments that are normally distributed.

We use the method of Wood and Chan [18] to generate an fBm field with $H = 0.6$ and 2^{17} points. We apply the fast, self-consistent density estimation method described in Section 2 to estimate the PDF of increments at distances of 2^1 to 2^9 grid points, with distance intervals that are integer powers of two. Figure 1c shows the PDF estimates of the standardized increments $\hat{f}(\Delta_x F)$. The standardized increment PDFs (colored curves) overlap strongly and are consistent with a normal distribution with zero mean and unit variance (the thick grey curve). The inset of Figure 1c shows that the moments of the PDFs scale as power laws (e.g., straight lines given log-log axes) of the increment distance. The structure functions are well-described by power laws as expected from Equation 4. We estimate the exponents of the structure functions using the York et al. [19] maximum likelihood method in log-log space, and we show in Figure 1d that the exponents vary as $H_m = 0.6 \cdot m$ as expected for an fBm field with $H = 0.6$ [17].

As noted at the beginning of this section, the goal of this analysis is to show whether (1) increments are normally distributed, and (2) the moments of the increment PDFs scale as a power-law of increment distance. This analysis technique uses the fast, self-consistent density estimation method as an efficient way of verifying that an fBm field has these characteristics. The standardized PDFs overlap and are all consistent with a normal distribution, which provides evidence that the increments are distributed normally. The approximate linearity of the structure functions in the log-log inset of Figure 1c provides evidence that the increment PDFs scale as a power-law of increment distance. And finally, the linearity of the H_m vs m points shown in Figure 1d provides further evidence that the increment PDFs are normally distributed. It can be shown that the moments of the normal distribution follow the relationship $\int_{-\infty}^{\infty} |x|^m \mathcal{N}(x)\, dx \sim \sigma^m$, where σ is the width of the normal distribution. From Equation 5, the PDF width is $\sigma = \sigma_o \cdot \Delta x^H$, so the moments should follow the relationship $M_m \sim \sigma^m \sim \Delta x^{H \cdot m}$. Therefore the $H_m = H \cdot m$ relationship demonstrated in Figure 1d is consistent with increments that are normally distributed and have PDF widths that vary as Δx^H.

4 Application to atmospheric model output

In a forthcoming manuscript (O'Brien, T. A., W. D. Collins, S. A. Rauscher, T. D. Ringler, M. Martini, W. Gustafson, and P. Ullrich, Fractal velocity fields cause resolution dependent updrafts in variable resolution atmospheric models. *Journal of Geophysical Research. In Prep.*), we develop a theory relating the distribution of vertical velocities (updrafts) in an incompressible atmospheric model to the probability distribution of horizontal velocity increments. We show that this theory predicts a resolution-dependent broadening of the vertical velocity distribution in a variable-resolution atmospheric model. In particular, for a self-similar horizontal velocity field with normally distributed horizontal increments, the theory predicts that the mean magnitude of vertical velocities $\langle |w| \rangle$ is simply related to the grid spacing Δx by $\langle |w| \rangle \sim \Delta x^{H-1}$, where H is the Hurst exponent that characterizes the self-similarity of the horizontal velocity field.

Our analysis of model output shows that the vertical velocity distribution broadens consistent with this Δx^{H-1} relationship. However, we have no a priori reason to expect that the horizontal velocity increments are distributed normally, and so it is unclear whether the observed broadening of the vertical velocity is truly consistent with our prediction. In order to characterize the distribution of horizontal velocity increments to evaluate this finding, it is necessary to estimate the PDF of $\mathcal{O}(10^5)$ sets of $\mathcal{O}(10^6)$ increment values. Given the amount of data reduction required in our analysis, and in fact in many other applications, a suitable method for estimating the PDFs should be as fast as possible to minimize the computational cost. The nuFFT-based improvement introduced in Section 2 reduces the computational cost of the BP11 method from approximately 10^2 seconds per estimated PDF to 1 second per estimated PDF (when applied to 10^6 data points). This reduces the computational cost of our analysis from $\mathcal{O}(10^3)$ CPU hours (e.g., a month on a serial processor) to $\mathcal{O}(10)$ CPU hours.

We apply the analysis presented in Section 3 to output from an atmospheric model with an idealized setup. We use output from the Community Atmosphere Model 4 (CAM4) [20], which is a modular hydrostatic atmospheric model with a variety of parameterizations that simulate various processes important for atmospheric dynamics (e.g., radiative transfer, convection, precipitation, etc.). We use a version of CAM4 that includes the Model for Prediction Across Scales atmospheric (MPAS-A) dynamical core, which predicts the evolution of the atmosphere by evalu-

ating conservation laws (e.g., conservation of mass, momentum, etc.) on a centroidal Voronoi tesselation of the sphere [14, 21].

The MPAS-A dynamical core is capable of operating on nonuniform grids that can effectively zoom in on an area of interest, which is one of the model's distinguishing features. Initial evaluation of CAM4 with the MPAS-A dynamical core showed that the model exhibits some distinctly resolution-dependent artifacts [14, 22]. Subsequent analysis has shown that this resolution-dependence may be linked to resolution dependence of the vertical velocity field [23], and we have recently developed a theory linking the resolution dependence of the vertical velocity field to the self-similarity of the horizontal velocity field. The theory relates the PDF of vertical velocities to the PDF of horizontal velocity increments.

To characterize the distribution of horizontal velocity increments at the model's highest resolution, we use the uniform-resolution 30km simulation described by Rauscher et al. [14]. We use one year of model output that is recorded for every 6 model hours. To facilitate this analysis, we have interpolated the CAM4 output from the its native, unstructured grid to a grid with uniform latitudinal and longitudinal spacing that has approximately the same 30km resolution as the native grid; in this grid, the globe is divided into 768 latitudes and 1,152 longitudes. The model is configured in accord with the aquaplanet protocol specified by Neale and Hoskins [24], in which the surface of the simulated planet is covered with water, and all boundary conditions are specified with rotational (in the direction of planetary rotation) and hemispheric symmetry. We leverage the rotational symmetry by treating latitudinal bands at a given level (altitude) as statistically identical, which we use to improve our sampling statistics.

At each time, latitude, and level in the model output, we calculate zonal velocity increments in the zonal direction (i.e., $\Delta_x U$, where U is the zonal wind velocity and x is the distance in the zonal direction). We calculate increments at all grid spacings that are powers of 2 between 2^0 to 2^{10}. We use the FFT-based density estimation method described in Section 2 to calculate the empirical characteristic function for each set of increments.

We parallelized the algorithm described in Section 2 by performing steps (1–5) in parallel for each time slice. We perform an additional step (6_0), in which we add the empirical characteristic functions from each time slice (treating values at each specific latitude and level separately) to obtain the empirical characteristic function for zonal velocity increments for the full year of model output. We then apply steps (6) and (8) to obtain the estimate of the probability density of zonal velocity increments for each latitude and level. We use these probability densities to estimate the 1^{st} through 9^{th} absolute moments of each distribution, which yield the 1^{st} through 9^{th} order structure functions of the zonal velocity field (see Section 3.2).

Figures 2 a, d, and g show the estimated probability densities of the zonal velocity increments $\hat{f}(\Delta_x U)$ from three distinct regions of the atmosphere: 40^oS at the 700 hPa level (approximately 3 km altitude), 0^oN at the 510 hPa level (approximately 5 km altitude), and 30^oN at the 970 hPa level (approximately 400 m altitude). These increment probability distributions, which are all standardized, overlap relatively well, which is consistent with self-similar behavior. Figures 2 b, e, and h show the first absolute moment of the increment distributions as a function of increment distance (i.e., the first-order structure functions). In all three figures, the first order structure functions exhibit approximate power-law scaling over a relatively wide range of increment distances, which is also consistent with self-similar behavior. The dashed gray lines in the figures show a power-law fit, using the York et al. [19] maximum likelihood method, to the structure functions for increment distances ranging between approximately 100 km and 500 km. We choose these bounds for two separate reasons. For the lower bound, it is well known that the diffusive properties of atmospheric models tend to dampen variability for length scales ranging from one grid cell to ten grid cells [13]. This effect manifests as a steepening of the first order structure functions for the two smallest increment distances (distances corresponding to 1 and 2 grid cells), so we restrict the fit to increment distances that are greater than or equal to 4 grid cells, which is approximately 100 km. Additionally, since it is hypothesized that there should be a scale-break for distances greater than approximately 500 km (e.g., [12]), we restrict our fit to increment distances less than or equal to this value.

We perform a similar power-law fitting procedure for the 1^{st} through 9^{th} order structure functions. Figures 2 c, f, and i show the estimated power-law slopes (the structure function exponents) H_m as a function of structure function order m. As discussed in 3.2, a self-similar field with normally-distributed increments should have structure function exponents that scale linearly with the structure function order, i.e.,: $H_m = H_1 \cdot m$. Such monofractal scaling is shown as a solid gray line in Figures 2 c, f, and i. The zonal velocity structure function exponents approximately follow this monofractal scaling for the 1^{st} and 2^{nd} order structure functions, but they diverge rapidly for the higher order exponents. This divergence is characteristic of a multifractal field [16], and it indicates that the zonal velocity increments are not distributed normally.

It is also clear from comparing the estimated distributions in Figures 2 a, d, and g with that of a unit normal distribution (shown as a gray dashed curve in all three figures) that the estimated distributions do not overlap well with the normal distribution. In exploring other potential distributions, we found that the increment distributions closely matched a standard logistic distribution–$f(x) \sim sech(x/2)$–over a wide portion of the atmosphere (shown as a solid gray curve in all three figures). However, it is apparent in Figure 2g that some areas of the atmosphere have zonal increment distributions that are quite positively skewed and are therefore inconsistent with symmetric distributions like the logistic distribution.

To demonstrate that the scaling properties of the zonal velocity field vary throughout the atmosphere, Figures 3 a and b show latitude-versus-height maps of H_1, and the excess kurtosis of the increment PDFs. (It is conventional in atmospheric sciences to express heights in terms of atmospheric pressure, which

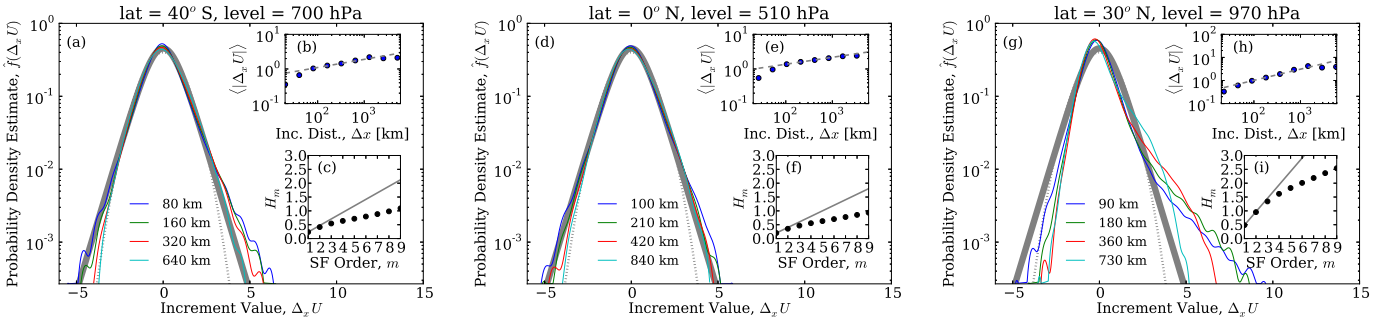

Figure 2: (a,d,g): Standardized zonal velocity increment distributions at various distances compared with a normal distribution (dotted gray line) and a logistic distribution (solid gray line); (b,e,h): their first order structure functions (blue dots) compared with a power law fit (dashed gray line); and (c,f,i): estimated exponents for the 1^{st} through 9^{th} order structure functions compared with the monofractal $H_m = m \cdot H_1$ relationship (solid gray line). Subfigures (a–c), (d–f), and (g–i) are grouped by location (see Figure 3, which marks these as locations 1, 2, and 3 respectively). Their locations are, respectively: 40^oS at the 700 hPa level (approximately 3 km altitude), 0^oN at the 510 hPa level (approximately 5 km altitude), and 30^oN at the 970 hPa level (approximately 400 m altitude).

decreases monotonically with height.) We calculate the excess kurtosis, γ_2 as follows:

$$\gamma_2 = \frac{\langle |\Delta_x U|^4 \rangle}{\langle |\Delta_x U|^2 \rangle^2} - 3, \tag{6}$$

and we average the excess kurtosis from increment distributions with increment distances ranging from approximately 100 km to 500 km (the same range as used in the power-law fit described previously).

Figure 3a shows that the first order structure function H_1 varies systematically throughout the atmosphere, with relatively small values near 0^o latitude and relatively large values near 40^o N/S. This is consistent with the first-order structure function of the water vapor field reported by Pressel [25] for a similar model configuration. It shows that the (modeled) atmosphere is not well-characterized by a single scaling exponent, as suggested by Nastrom and Gage [12], but that the fractal behavior of the atmosphere ranges from anti-persistent ($H_1 < 0.5$) to persistent ($H_1 > 0.5$) depending on location.

Further, the excess kurtosis, γ_2, which is a parameter that describes the 'peakedness' of a distribution relative to the normal distribution, also varies throughout the atmosphere. Figure 3b shows that γ_2 varies from approximately 1 at 0^o latitude to approximately 7 near 30^o N/S. A normal distribution is characterized by zero excess kurtosis, whereas distributions with sharper peaks and fatter tails (relative to the normal distribution) have positive excess kurtosis. The logistic distribution has $\gamma_2 = 1.2$, which is consistent with values over a wide area of the equator.

Interestingly there are zones of high kurtosis near 1000 mb at approximately 30^o N/S; these leptokurtic zones are associated with positive skew. Examination of Figure 2g, which shows the estimated increment distributions from this high-kurtosis zone, reveals that the negative half of the distribution overlaps reasonably well with the normal distribution, whereas the positive half of the distribution has wide tails. This positively skewed distribution reflects an abundance of zones in which the wind speed tends to accelerate in the eastward direction, which is indicative of a force acting in that direction. That this skewed distribution occurs in the midlatitudes (near 30^o latitude), where the effect of Earth's rotation becomes important, suggests that the Coriolis force may be the cause of the skewed distribution.

5 Discussion

5.1 Improving the speed of ECF-based methods using the FFT

While we could have used other methods of density estimation, such as binning or traditional kernel density estimation, the Bernacchia and Pigolotti [1] method avoids the complication of having to choose either bin width or kernel bandwidth, which is a subjective choice when faced with data from an unknown distribution. The Bernacchia and Pigolotti [1] method simultaneously and objectively determines both the optimal shape and optimal bandwidth for a kernel density estimate. However, because the Bernacchia and Pigolotti [1] method involves a transformation from data-space to Fourier-space (i.e., calculation of the empirical characteristic function), the method is quite slow if the empirical characteristic function is calculated using a direct Fourier transform. We show in Sections 2 and 3 that replacing the direct Fourier transform with a nonuniform FFT can dramatically increase the speed of the method without compromising the accuracy of the method.

As far as we are aware, no authors have explored the use of nonuniform FFT methods for calculating empirical characteristic functions (ECFs) in general, even though empirical characteristic functions have a wide variety of uses, including: testing for distribution symmetry [26], testing for data independence [27, 28], testing whether data belong to a given distribution family [6, 29], testing whether two sets of data belong to the same family [30],

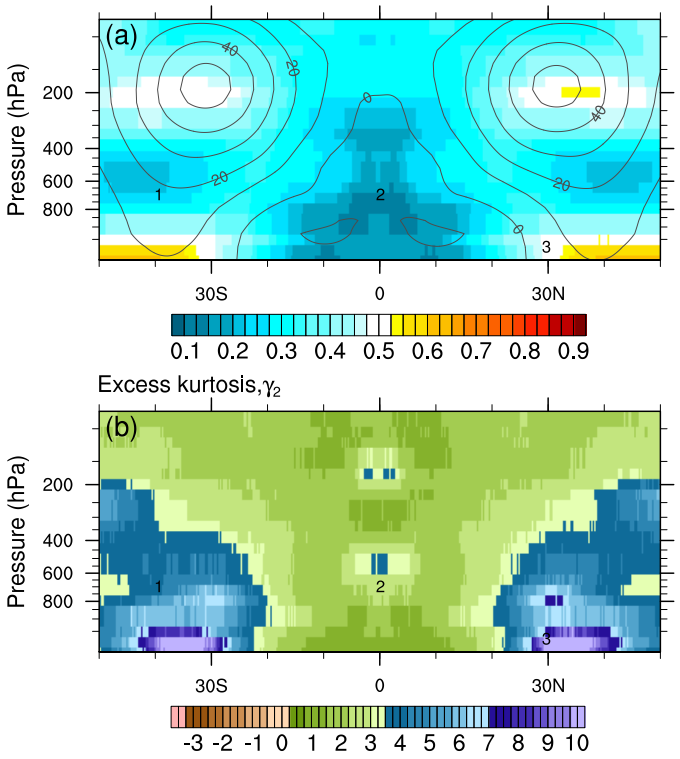

Figure 3: (a): A latitude vs. height map of the estimated power law exponent from the first order structure function of the zonal velocity increments. The gray contours depict lines of constant zonal velocity (m/s) to show the location of the midlatitude jet streams. (b) A latitude vs. height map of the excess kurtosis (see text for definition) of the increment PDFs; for reference, a normal distribution has zero excess kurtosis. In both figures, the black numbers 1, 2, and 3 indicate the locations, respectively, from which Figures 2 (a–c), (d–f), and (g–i) are calculated.

and model fitting [31, 32, 33]. While the results in Figure 1b show timings for the full BP11 density estimation, the computational time of the method is dominated by calculation of the ECF. The results in Figure 1 (and the order-of-magnitude calculations presented in Section 2.3) show that the nonuniform FFT can increase the calculation speed of the ECF by two orders of magnitude while keeping the approximated ECF accurate to the 7^{th} decimal place.

The nonuniform FFT could be especially beneficial for calculating the ECF of multidimensional datasets. In this manuscript, our use of the nonuniform FFT method is limited to one-dimensional data, since the BP11 density estimation has so far only been developed for univariate data. However, the nonuniform FFT method is also applicable to multidimensional transforms [9], and so the idea developed in this manuscript could easily be extended to multidimensional data. For N sets of d-dimensional data, direct calculation of the ECF onto a Fourier grid with M frequencies in each dimension requires $\mathcal{O}(d \cdot N \cdot M^d)$ calculations. On the other hand, a nonuniform FFT method that uses a q-point convolution requires $\mathcal{O}\left(N \cdot q^d + M^d \cdot log(M^d)\right)$ calculations. Following the same assumptions in Section 2.3 ($q < M$ and $logM \ll N$) the nonuniform FFT method is roughly $\mathcal{O}\left((M/q)^d\right)$ times faster than the direct calculation. Both the direct and nuFFT ECF calculation methods suffer a 'curse of dimensionality' (i.e., the computational complexity scales as a power of the dimensionality), but the nonuniform FFT method reduces the negative impact of increased dimensionality by only applying the convolution to a relatively small q^d hypercube of points surrounding each datum. For $M/q \sim \mathcal{O}(100)$, as in this manuscript, an FFT-based calculation of the ECF for bivariate data would be $\mathcal{O}(10,000)$ times faster than the direct calculation.

5.2 *Summary*

The analysis in Section 4 shows that the zonal velocity increments in our atmospheric model output have increments that are clearly not distributed normally (Figures 2 a, d, and g) and that the field is multifractal (Figures 2 c, f, and i). Based on the excess kurtosis values shown in Figure 3b, no portion of the model's atmosphere has increments that are normally distributed. This analysis has shown that our theory relating the self-similarity of the horizontal wind field to the distribution of vertical velocities, which was developed based on a wind field with normally distributed increments, needs to be generalized to account for a broader range of distributions. The ability to rapidly and robustly characterize the zonal velocity increment distributions has thus proved invaluable for helping us advance our scientific work.

This manuscript generally shows that nonuniform FFT methods can be used to dramatically reduce the computational cost of the empirical characteristic function. Though this manuscript focuses specifically on the case of using the nonuniform FFT to improve the ECF calculation stage of the Bernacchia and Pigolotti [1] estimation method, this method should be applicable to other ECF-based methods. We posit that the nonuniform FFT would especially reduce the computational cost of multidimensional ECF calculations: potentially by a factor of $\mathcal{O}(100^d)$ for d-dimensional data.

If the BP11 method can be extended to multidimensional data, then a nonuniform FFT method could be used to dramatically decrease the computational time of the method. Combined with the nonuniform FFT, a multidimensional BP11 method could provide an objective, fast, and robust way to estimate multivariate probability distributions. For the purposes of atmospheric research, such a method could be invaluable for characterizing the interdependency of atmospheric state variables. For example, a multidimensional BP11 estimate of the joint velocity, humidity, and enthalpy PDF could provide a non-parametric method for estimating subgrid fluxes that is complementary to existing parametric methods (i.e., [34]), which are known to depend on the shape of the assumed PDF [35]. While Bernacchia and Pigolotti [1] suggest that their method readily extends to multiple dimen-

sions, special care will be required to develop multidimensional frequency filters (i.e., I_n in Section 2.3), since neither the filter used in this paper nor the filter used by Bernacchia and Pigolotti [1] have simple multidimensional analogs.

The BP11 self-consistent density estimation method is an objective and robust way to estimate the underlying distribution of univariate data. As we show in this manuscript, use of a nonuniform FFT can reduce the computational cost of the method by a factor of approximately 100. This modification makes the BP11 method fast relative to human timescales: it requires less than a second to estimate the PDF of 10^5 data points using Python code on a modern PC. This makes the FFT-based BP11 method a viable alternative to histogram-based methods in data analysis software (e.g., SciPy or R). Toward this goal, the lead author is working with his home institution to release the code used in this manuscript under a free (e.g., GNU) license, so that he may pursue including it in the SciPy *stats* package.

A Definition of terms

velocity increment: the difference between two points in a field at a given distance: $\Delta_x F \equiv F(x) - F(x+\Delta x)$.
structure function, m^{th}-order: the m^{th} moment of the increment distribution, as a function of increment distance: $\langle |\Delta_x F|^m \rangle$.
resolution: the physical size of a grid element in an atmospheric model. Unless otherwise specified, this typically refers to the horizontal size of grid elements.
vertical velocity: the velocity of air in the direction perpendicular to the Earth's surface.
horizontal velocity: the velocity of air in the directions parallel to the Earth's surface.
zonal velocity: horizontal velocity in the direction parallel to the direction of Earth's rotation (i.e., parallel to lines of latitude).
fractal velocity field: a velocity field that is statistically self-similar. The statistical distribution of velocities has the same basic form regardless of the physical scale at which the distribution is calculated, and whose width is a power-law of the physical scale. The structure functions of such a field are power laws of the increment distance.
Hurst exponent, H: an exponent that characterizes the properties of a monofractal field.
monofractal field: a field in which the increments are normally distributed. The moments of the field are given by $\langle |\Delta_x F|^m \rangle \sim \Delta x^{H \cdot m}$.
multifractal field: a field in which the increments are not normally distributed. The moments of the field are given by $\langle |\Delta_x F|^m \rangle \sim \Delta x^{\mathscr{H}(m)}$, where $\mathscr{H}(m)$ is a non-linear function of the structure function order.

Acknowledgments

The authors would like to thank two anonymous reviewers for comments that helped improve the quality of the manuscript. The authors would additionally like to thank Prof. Christopher Paciorek of UC Berkeley for comments that helped to focus the discussion in this manuscript and Dr. Karthik Kashinath of Lawrence Berkeley National Lab for helping to improve the mathematical presentation of the manuscript.

This research was supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy Regional and Global Climate Modeling Program (RGCM) and used resources of the National Energy Research Scientific Computing Center (NERSC), also supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

[1] A. Bernacchia, S. Pigolotti, Self-consistent method for density estimation, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (3) (2011) 407–422, ISSN 1467-9868, doi:10.1111/j.1467-9868.2011.00772.x, URL `http://dx.doi.org/10.1111/j.1467-9868.2011.00772.x`.

[2] B. W. Silverman, Density estimation for statistics and data analysis, vol. 26, CRC press, 1986.

[3] D. S. Wilks, Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2006.

[4] R. Srihera, W. Stute, Kernel adjusted density estimation, Statistics & Probability Letters 81 (5) (2011) 571–579, doi:10.1016/j.spl.2011.01.013.

[5] M. Wand, M. Jones, Kernel Smoothing. Vol. 60 of Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1995.

[6] K. Murota, K. Takeuchi, The studentized empirical characteristic function and its application to test for the shape of distribution, Biometrika 68 (1) (1981)

55–65, URL http://biomet.oxfordjournals.org/content/68/1/55.short.

[7] G. S. Watson, M. Leadbetter, On the estimation of the probability density, The Annals of Mathematical Statistics 34 (2) (1963) 480–491.

[8] J. W. Cooley, J. W. Tukey, An algorithm for the machine calculation of complex Fourier series, Mathematics of computation 19 (90) (1965) 297–301.

[9] L. Greengard, J.-Y. Lee, Accelerating the nonuniform fast Fourier transform, SIAM review 46 (3) (2004) 443–454, doi:10.1137/S003614450343200X, URL http://epubs.siam.org/doi/abs/10.1137/S003614450343200X.

[10] J. Liao, Y. Wu, Y. Lin, Improving Sheather and Jones bandwidth selector for difficult densities in kernel density estimation, Journal of Nonparametric Statistics 22 (1) (2010) 105–114, URL http://www.tandfonline.com/doi/abs/10.1080/10485250903194003.

[11] A. Dutt, V. Rokhlin, Fast Fourier transforms for nonequispaced data, SIAM Journal on Scientific computing 14 (6) (1993) 1368–1393.

[12] G. Nastrom, K. Gage, A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft, J. Atmos. Sci. 42 (9) (1985) 950–960.

[13] W. C. Skamarock, Evaluating Mesoscale NWP Models Using Kinetic Energy Spectra, Mon. Wea. Rev. 132 (12) (2004) 3019–3032, ISSN 0027-0644, doi:10.1175/MWR2830.1, URL http://dx.doi.org/10.1175/MWR2830.1.

[14] S. A. Rauscher, T. D. Ringler, W. C. Skamarock, A. A. Mirin, Exploring a Global Multi-Resolution Modeling Approach Using Aquaplanet Simulations, J. Climate *In Press*, ISSN 0894-8755, doi:10.1175/JCLI-D-12-00154.1, URL http://dx.doi.org/10.1175/JCLI-D-12-00154.1.

[15] B. B. Mandelbrot, The fractal geometry of nature, Times Books, 1983.

[16] A. Davis, A. Marshak, W. Wiscombe, R. Cahalan, Multifractal characterizations of nonstationarity and intermittency in geophysical fields: Observed, retrieved, or simulated, J. Geophys. Res. 99 (D4) (1994) 8055–8072, ISSN 2156-2202, doi:10.1029/94JD00219, URL http://dx.doi.org/10.1029/94JD00219.

[17] A. Davis, A. Marshak, W. Wiscombe, R. Cahalan, Multifractal characterizations of intermittency in nonstationary geophysical signals and fields, in: G. Treviño, J. Hardin, B. Douglas, E. Andreas (Eds.), Current Topics in Nonstationary Analysis. Proceedings of the Second Workshop on Nonstationary Random Processes and their Applications. San Diego, California 11-12 June 1995, CHIRES ASSOCIATES INC HOUGHTON MI, 1996.

[18] A. T. A. Wood, G. Chan, Simulation of Stationary Gaussian Processes in [0, 1] d, Journal of Computational and Graphical Statistics 3 (4) (1994) 409–432, ISSN 1061-8600, doi:10.1080/10618600.1994.10474655, URL http://dx.doi.org/10.1080/10618600.1994.10474655.

[19] D. York, N. M. Evensen, M. L. Martínez, J. De Basabe Delgado, Unified equations for the slope, intercept, and standard errors of the best straight line, American Journal of Physics 72 (3) (2004) 367–375, doi:10.1119/1.1632486, URL http://scitation.aip.org/content/aapt/journal/ajp/72/3/10.1119/1.1632486.

[20] R. Neale, J. Richter, A. Conley, S. Park, P. Lauritzen, A. Gettelman, D. Williamson, S. Vavrus, M. Taylor, W. Collins, et al., Description of the NCAR Community Atmosphere Model (CAM 4.0), NCAR Technical Note, National Center of Atmospheric Research TN-485+STR.

[21] W. C. Skamarock, J. B. Klemp, M. G. Duda, L. D. Fowler, S.-H. Park, T. D. Ringler, A Multiscale Nonhydrostatic Atmospheric Model Using Centroidal Voronoi Tesselations and C-Grid Staggering, Mon. Wea. Rev. 140 (9) (2012) 3090–3105, ISSN 0027-0644, URL http://dx.doi.org/10.1175/MWR-D-11-00215.1.

[22] T. A. O'Brien, F. Li, W. D. Collins, S. A. Rauscher, T. D. Ringler, M. Taylor, S. M. Hagos, L. R. Leung, Observed Scaling in Clouds and Precipitation and Scale Incognizance in Regional to Global Atmospheric Models, J. Climate 26 (23) (2013) 9313–9333, ISSN 0894-8755, doi:10.1175/JCLI-D-13-00005.1, URL http://dx.doi.org/10.1175/JCLI-D-13-00005.1.

[23] Q. Yang, L. Ruby Leung, S. A. Rauscher, T. D. Ringler, M. A. Taylor, Atmospheric moisture budget and spatial resolution dependence of precipitation extremes in aqua-planet simulations, J. Climate ISSN 0894-8755, doi:10.1175/JCLI-D-13-00468.1, URL http://dx.doi.org/10.1175/JCLI-D-13-00468.1.

[24] R. B. Neale, B. J. Hoskins, A standard test for AGCMs including their physical parametrizations: I: The proposal, Atmosph. Sci. Lett. 1 (2) (2000) 101–107, ISSN 1530-261X, doi:10.1006/asle.2000.0022, URL http://dx.doi.org/10.1006/asle.2000.0022.

[25] K. G. Pressel, Water Vapor Variability Across Spatial Scales: Insights for Theory, Parameterization, and Model Assessment, Ph.D. thesis, University of California, Berkeley, 2012.

[26] A. Feuerverger, R. A. Mureika, The empirical characteristic function and its applications, The annals of Statistics

5 (1977) 88–97, URL http://www.jstor.org/stable/10.2307/2958763.

[27] S. Csörgő, Testing for independence by the empirical characteristic function, Journal of Multivariate Analysis 16 (3) (1985) 290–299.

[28] M. Bilodeau, P. Lafaye de Micheaux, A multivariate empirical characteristic function test of independence with normal marginals, Journal of Multivariate Analysis 95 (2) (2005) 345–369.

[29] L. Baringhaus, P. D. N. Henze, A consistent test for multivariate normality based on the empirical characteristic function, Metrika 35 (1) (1988) 339–348, URL http://link.springer.com/article/10.1007/BF02613322.

[30] V. Alba Fernández, J. Gamero, J. Muñoz Garcia, A test for the two-sample problem based on empirical characteristic functions, Computational statistics & data analysis 52 (7) (2008) 3730–3748, URL http://www.sciencedirect.com/science/article/pii/S0167947307004847.

[31] Y. Fan, Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function, Journal of Multivariate Analysis 62 (1) (1997) 36–63.

[32] J. L. Knight, J. Yu, Empirical characteristic function in time series estimation, Econometric Theory 18 (03) (2002) 691–721, URL http://journals.cambridge.org/abstract_S026646660218306X.

[33] M.-D. Jiménez-Gamero, V. Alba-Fernández, J. Muñoz-García, Y. Chalco-Cano, Goodness-of-fit tests based on empirical characteristic functions, Computational Statistics & Data Analysis 53 (12) (2009) 3957–3971, URL http://www.sciencedirect.com/science/article/pii/S0167947309002308.

[34] J.-C. Golaz, V. E. Larson, W. R. Cotton, A PDF-based model for boundary layer clouds. Part I: Method and model description, Journal of the atmospheric sciences 59 (24) (2002) 3540–3551, URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0469(2002)059%3C3540:APBMFB%3E2.0.CO%3B2.

[35] P. A. Bogenschutz, S. K. Krueger, M. Khairoutdinov, Assumed Probability Density Functions for shallow and deep convection, Journal of Advances in Modeling Earth Systems 2 (4).